DATE: 14-2-2017 SUBJECT: داسثريبة

Big data ⇒ refers of exponential growth and availability of data

"Volume" large size - more complex tools - variety - volicity - value or not "[illegible]" - Trusted or not "accurcy of data".

data structure " structured - semi structured - unstructured"

structured	semi structured	unstructured
relational dB	XML	video
	Jason	audio
	Binary Jason	PdF

أجزاء مش ليه organize

required tools :

- way of storage
- analysis Tools
- data scientist → personal scales "مثال" → technical knowledge → mathematics tools

oLaP ⇒ online analatical processing

"التطور والتحرير" هام

Slide 17
Island
seperable
"spreadsheets"

1- not secure
2. replication
3- hacking
4-

1- easy and simple
2-

data ware house.

الغرض الأساسي
هو تحرير البيانات من الآخر أمان

1- Controlled
2- must have permission for apply any action on data
3- controlled by data Base administrator.

SandBox
permission يتم عمل dWH ثم أخذ جزء منها
وبعد عمليات على هذا الجزء إلى حين احتياجه من dWH فقط وتنفيذ
عملية العمليات المطلوبة

DATE : SUBJECT :

وبذلك يبقى Highperformance

Case study. Bank from local int global

١- القوانين الخاصة بكل دولة.

٢. طرق التعامل مع البنك.

٣- High security.

٤- تغير العملات.

٥- بزيادة العملاء نحتاج إلى عالة زيادة.

٦- ممكن نحتاج تعمل distributed dB.

Data scientist 1_ variety.

2_ stream large.

3_ managment data.

4_

• Data analist

• بيشوف data محتاجه إيه وليا عملناها تعمل لوحدها

KPI ⟶ Keep performance indicator.

Classifier ⟶ TP , FN , TN , FP

performance evaluation.

*Bussines intelligance بس بنعمل هو منطلقه المعايير وهو بيقسم بيانات

ويعمل analysis علشان أقدر أخد قرار بناء على هذا التحليل.

مش بيقدر يتعامل إلا مع "structured data" فقط.

DATE: ______ SUBJECT: ______

. Data scientist.

آلترجبره ولماذا ده من السابقه

« unstructured - structured » data دعقد ، يتعامل مع أى نوع من

Comparison between them.	① scales - ② Job of each of them } هام

33

Throughput => الذى أيه حيث تتلف

البيانات من شيء لآخر data

Skales of data scientist lec2

1- statistices skills
2. Database
3- Critical thinking, creative, Innorate + Communications skills + Adaptive
4- machine learning + Data mining + Advanced mathematics.

5- collect data from different online source
6- Extract data & Analysis
7- can make Correlations & Connections.
8- web development & web design.
9- programming skills,

S5

Data enables. ≃ data collectors.

Professional Business reports, machine learning limit tools

DATE: 57

- Quantitive — أي بيقدر يقيس الداتا الي جاية ويعمل statistics ويعرف يعمل Technical report

- Skeptical — تيكافئ الي يكون شكاك

- Communications & Collaborative skills — لديه القدرة على التعامل مع الناس

S.10

الأشياء دي مش هتأتي — pandemics أوبئة

S12 — Swine Flu انفلونزا الخنازير

outcome — بدأ يشوف علاج المرض بـ 5 مراحل بتوعه

S13 Life science. "Genome" الصفات الوراثية

A1. 5 , Complexity.

A2.

life cycle of DAP

lec3

1- define problem
1- outcomes
2- I/P
3- issuse.
4- resources available

2- collect avalable data and besure that it is enough and secure.

3- modeling.

4- Test model.

- يفضل بعد الانتهاء من كل مرحلة التأكد أنه تم تطبيقها بأفضل شكل ممكن.

documentation.

DATE: ____________ SUBJECT: ____________

مع كل مرحلة يتم التنبؤ لكن يتم تفسيرها الآخر بعد ذلك.
وأيضاً ايه اي عمل additional analysis.
والتأكد من validation للداتا المختلفة

Business user → end user.

S. 9

① Discovery.

- learn aboute the problem domain.
- hold history of this domain and analysis it
- Documentation of the old project.
- measure to what resource available and period of time and Quality of data.

2- Data pre
is responsable of building "Sand Box".

3. model
1- Sw that will uses.
2- Feature important "Feathure selection"
3- HW.

S.16
probe. إستجوب

S 20 discussion in next lecture.

DATE: _____ SUBJECT: _____

Big data ⇒ is a popular term which refers to the exponential growth and availability of data, both structured and unstructured.

Three Vs' to describe the definition of big data.

Volume
velocity
variety

Volume.

⇒ There has been a large increase of data volume there reasons.

1. All of the transactional data that has added up over the years.
2. streaming data From social media.
3. machine to machine data increase.

velocity.

⇒ Data is being streamed at huge speeds and needs To be dealt with a timely manner

1. Social media.
2. mobile devices.

the biggest challenge is how to react Fast enough. To the massive amount of data that is being Flew rapidly.

DATE: _____ SUBJECT: _____

Variety

managing all the different formats is an issue many organizations have to battle.

- there are many different types of data
- Structured.
- Email.
- Audio & video.
- Application data.
- Financial transactions.
- Unstructured documents.
- So to manage many organizations have to battle

Volume, velocity, variety, value, veracity.

Big Data

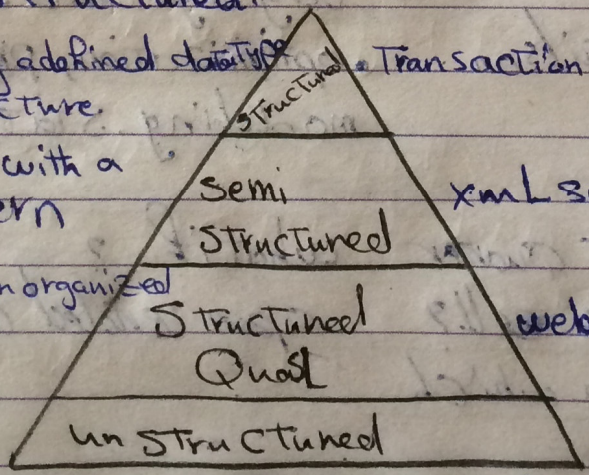
is data whose scale, distribution, diversity, and or timeliness require the use of technical architectures and analytics to enable.

Key charactristics of BD

1- data volume.

2- processing complexity ← parallel computing environment and massively parallel processing

3- Data Structured.



Data containing a defined data type / formate, structure. ← Transaction data and OLAP

Textual data files with a discernable pattern ← xml schema

Textual data and unorganized data format ← web click stream

ex Text document, pdf

Data has no inherent structure and in usually stored as different types of files.

DATE: _____ SUBJECT: _____

Storing data

Data islands spreadmarts

isolated data

- spreedsheet and Low volume DB.
- Analyst dependent on data extracts.

Data warehouse

Centeralized Data Containers in a purpose-built space.

- Analyst dependent on IT @ DBAs For data access and schema change.
- Analysts must spened significant time to get extract For multiple sources.

Analytic SandBox

Data assets gethered From multiple sources and technologys For analysis.

- enable high performance
- reduce cost associated with data replicated.
- analyist "owned"
- more robust analyses

Business inteligence

Structured data, traditional sources. manageable datasets.

Standard and details on demand.

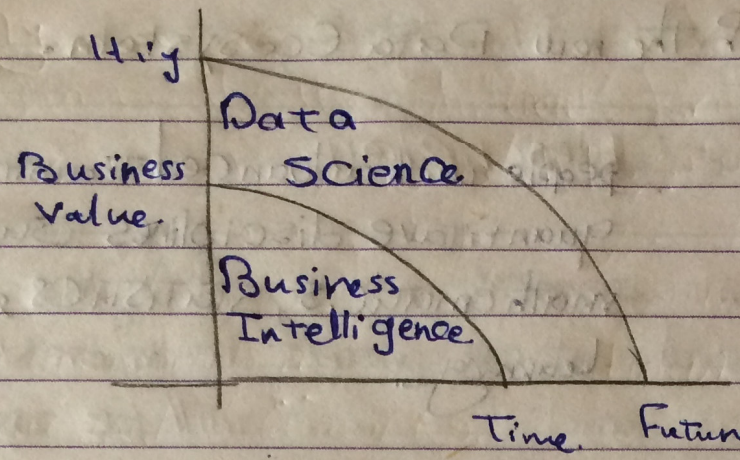
- what happened last quarter How many did we sell? where is the problem? in which situation.

Data science

- Structure/unstructured data, many types of sources, very large data sets.
- optimization, predictive modeling. Statistical analysis.

what if..?
open ended questions.

DATE: _____ SUBJECT: _____



implications of typical Architecure For data science.

1- high-value data is hard to reach and leverage
2. Data is moving in batches From EDW to local analytics.
3. isolated and analytic projects, rather than centrally-managed of analytics.

. the Big data trend is gererating an enormous amou of information that requires advanced analytics and new market players to take advantage of it.

Criteria For Big Data projects.
1- speed of decision making
2. throughput
3. Analysis Flexibility.

DATE: _____ SUBJECT: lec 2

Three Keys roles of the new Data Ecosystem.

- Deep analytical Talent (Data scientist) — people with advanced training in quantitave disciplines such as mathematics, statistics, machine learning.

- Data sawy professionals (Analysts, Data sawy managers) — people with abasic Knowledge of statistics and/or machine learning who can define key questions that can be answered using advanced analytics.

- Technology & Data Enablers — people providing technical expertise to support analytics projects skill sets including Computer programing and DB administrator.

Data scientist Key Activities

1- reFrame business challenges as analytics challenges
2- Design implement and deploy statistical models and data mining Techniques on big data.
3. Create insights that lead to actionable recommendations.

DATE: _____ SUBJECT: _____

«Data Analytics life cycle»

- Value of using the data Analytics lifecycle.

1- Ensure rigidity and completeness.
2- Enable better transition to members of the cross-Functional analytic teams...

Creating and documenting a process will help demonstrate rigor in your findings.

- repeatable.
- Scale to additional analysts.
- Support validity of findings.

Need for a process to Guide Data Science projects.

1- well-defined processes can help guide any analytic project
2- Focus of Data Analytics ~~project~~ lifecycle is on Data Science projects, not business intellegence.
3. Data science projects tend to require a more consultative approach, and differ from BI projects in a few ways.

- less predictable data
- more projects which lack shape or structure.
- more due diligence in discovery phase.

DATE: _____ SUBJECT: _____

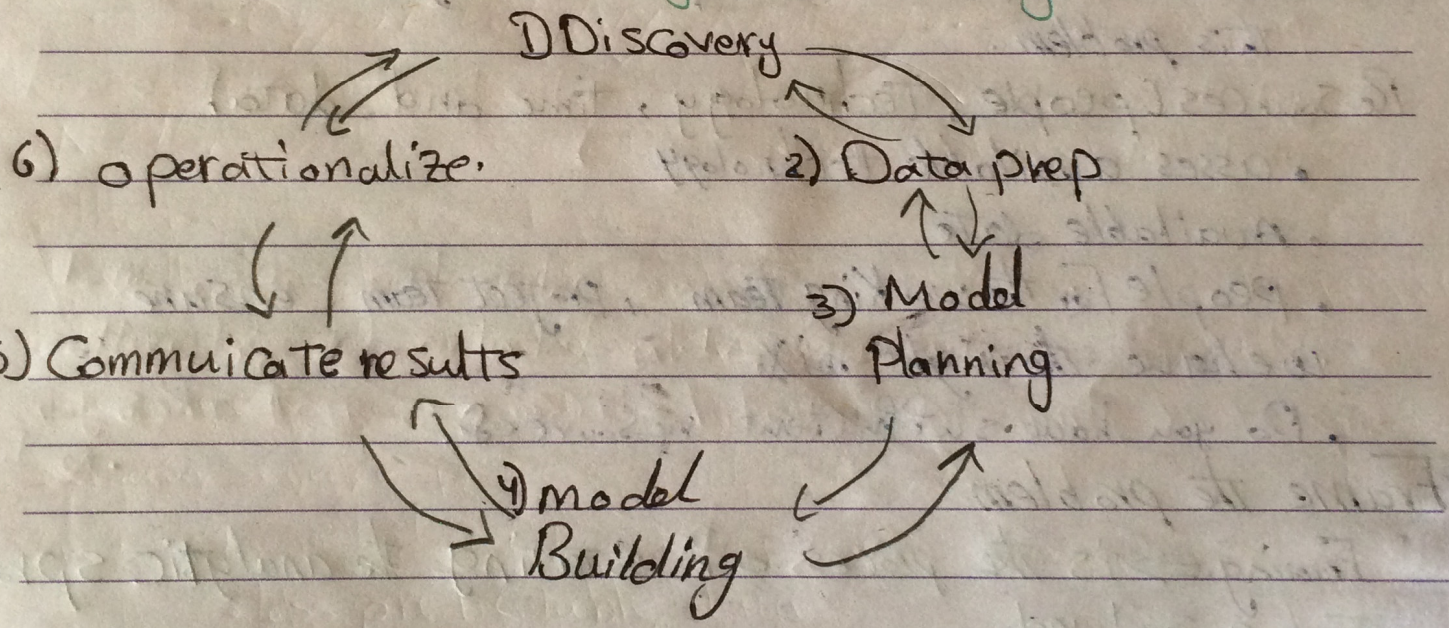
Key roles for a successful Analytic project.

Role	Description
Business user	Someone who benefits from the end results and can consult and advise project team on value of end results and how these will be operationalized
project sponsor	person responsible for the genesis of the project, providing the motives for the project and core business problem, generally provide the funding and will measure the degree of value from the final outputs of the working team.
project manager	Ensure key objectives are met on time and at expected quality.
Business intelligence Analyst	Business domain expertise with deep understanding of the data KPIs, key metrics and business intelligence from a reporting prespective.
Data Engineer	Deep technical skills to assist with tuning SQL queries for data management extraction and support data realize to analyti sandBox

DATE: _____ SUBJECT: _____

Database Administrator (DBA)	Database Administrator who provisions and configures database environment to support the analytical needs of working team
Data scientist	provide subject matter expertise for analytical techniques, data modeling, applying valid analytical techniques to given business problems and ensuring overall analytical objectives are met

Data Analytics Lifecycle



* Data Analytics Lifecycle :-

We can go back and refine work done in prior phases given new insights and information that you've uncovered

Phase 1: Discovery.

• learn the Business Domain.

determine amount of domain knowledge needed to orient you to data and interpret results down stream.

• Determine the general analytic problem type "such as clustering, classification"

• then conduct initial research to learn about the domain area you'll be analyzing.

• learn From the past.

• Learn how previous attempts in the organization to solve this problem.

Resources (people, technology, time and data)

• asses available technology

• Available data

• people for the working team, project team ensure we have the right mix.

• Do you have sufficient resources.

Frame the problem

Framing ⇒ is the process of stating the analytics problem to be solved.

• state the analytics problem, why it is important

Summary of discovery phase.

1) learn about domain knowledge.

2) detect available resources.

(1 زي امكانيات الاستخدام -Tools -data -technology ....)

تحديد نوع learning "clustering أو classification"

DATE: Lec 4 SUBJECT:

« data preperation phase »

تعتبر أطول وأصعب مرحلة.

حيث هنا بتكون هنا Sand Box وقد تكون بعض الأحيان البيانات كبيرة جدا data ware house.

١- بأخذ كل data الممكنة متاحة

٢- تكوين هنا Sand Box هنا بتعمل Limited data:

ويوجد هنا 2 technologies

1) ETL extract Transform load.
2) ELT extract load Transform.

« deep analysis » حيث يتجمع ويتعمل كل الداتا المتاحة وجميعها الى هو بايزه داخل DWH, it وهنا يتم التعامل مع

٣- check data.

Valid
Concistance
process
missing fields

1. Check data Type (structured, unstructure, semi structured),
2. systematic error

1- prepare Analytic SandBox
2- perform ELT
3- Familiarize youreself with the data thoroughly
4- Data Conditioning
5- Survey & visualize.
6- Determine methods
7- Techniques & workFlow.

DATE: ______ SUBJECT: ______

* phase 3 :- Model planning

1- Data exploration
2- Variable selection
3- Model Selection :

١. تحديد الـ Framework بتاعي و أيه الـ Techniq اللي هستخدمها للأداء.
٢. الـ workflow, Feature Selection اللي تحدد ومين الـ model هيخرج.
٣. تحديد الشكل النهائي لـ model بتاعنا.

* phase 4 :- model Building

1- develop data sets for testing, Training, and production purposes.
2- get the best environment you can for building models and workflows. (R, RPL, ...)

١. بنشوف نوع data وهل هتكون مناسبه ولا لا.
٢. التعديل في ملعب الـ model اللي أعكسه لتلد شيء أو بعض الأخطاء.
important parameters - Valid - accuracy, mistakes وأقل ما يمكن.

* phase 5 :- Communicate Results

١- أي محاولة اقناع Sponser بالقرارات التي تم تحديدها سابقاً.
٢. وعرض الشيء نخرج عليه.
Did we succeed? Did we fail?

DATE: _____ SUBJECT: _____

* phase 6 هو operationalize

• Test الـ system و check الـ pipeline.

4 Core Deliverables to Meet most Stakeholder Needs

1 - presentation for project spoasors

- Big picture takeaways for executive level Stakeholders.
- Determine key messages to aid their decision-making process
- Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp.

2 - presentation for Analysis

- Business process changes
- reporting changes
- Follow Data Scientists will want the details and are comfortable with technical graphs

3 - Code: For technical people.

4 - Technical speces & implementing the code.

⇒ Analyst wish list for a successful Analytics project.

* Data & workspaces.

1 - Access all data

2 - up-to-date data dictionary

3 - Ability to move data back between staging

4 - Sand box.

DATE: _____ SUBJECT: _____

Tools:.

- statistical, mathematical, visual SW.
- Tool or place to log errors with systems.
- Collaboration → online platform for communication with team members.

DATE: lec. SUBJECT:

estimated	Actual	
+ ve	+ ve	→ Tp
- ve	- ve	→ TN
+ ve	- ve	→ FP
- ve	+ ve	→ FN

$$Accuracy = \frac{TF + TN}{Tp + TN + FN + FP}$$

. which i/p
. Bulding model that solve the problem
. Check accuracy of model → Compare with myself
↳ Compare " others in same problem

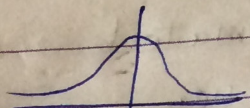
*hypotheis testing. وذلك بنفس نتائج دنفس samples يعني زي ما هي
↳ Choosing variable → لو حصل فرق في الأفراد الاختيار
↳ لو مش هيحصل فرق بيتغاضى عنه → وذلك عدم ر يعرف mean variance

العرض من بناء model هنا ليس تكوين نتائج ولكن من أجل تنبيه نتائج موجوده بالفعل.

*median ↓ الوردي يبقى المتصف.
زوجي بنجمع 2 في النص ونقسم على 2.

وجدوا عند حساب mean ان هناك احتمالية الحصول على 0
↳ هنا عدد نجد ان 2 زي بعضها في حين قد تكون ± زي مختلفين لذلك تم اللجوء الى variance

Test check difference of mean.



يشترط على normal distribution.

1. t. test الأصلي → default of R
2. welchs t. test → more tightly.
3. Rank sun → general << less tightly.
عامة استخدامه لأي نوع من الداتا.

H0 ⇒ "hypothes " وتعني مش هيحدث أي تغير
H1 ⇒ " " " تعني وجود حل جديد

T. test " two way test " الأصلي - inter
1- يشترط الـ 2 distribution مش ليهم نفس variance.

أما student T. test يشترط الـ 2 يكون ليهم نفس variance.

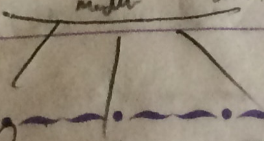
(1)
→ . t = 0 accept null hypotes. مفيش فرق
→ . t = w area under the curve. في فرق
accept intermediate hypothesis. ← ينتج منه

RackSun.

w ⟹ old new threshold value.

Samples			A	
X_1	X_2	X_3	X_{11}	X_{21}
			X_{12}	X_{22}

$w = \sum sgn(-old + new)$
model model



X →	0	+	−
sgn(x)	0	1	-1

DATE: ______ SUBJECT: ______

power ⟹ positive اكبر القيمة الي

signif ⟹ F P error قيمة

effect size ⟹ Actual difference between 2 means.

ANOVA

يقدر اشتغل على one way (1 variable) ولا Two way (2 way).

• مش شرط يكونو بجنب بعض 2 model , لها مين.

• و ممكن يكونو في أكثر من 2 model

528 مهم جدا

1. Calculate mean of each population.

$m_1 = 2.67$

$m_2 = 2.67$

$m_3 = 3$

①	②	③
1	2	2
2	4	3
5	2	4

2- grand mean $m_G = \frac{m_1 + m_2 + m_3}{n} = \frac{2.67 + 2.67 + 3}{9} = 2.78$

2- Sum of Squares (SS)

$SS_{within} = \sum (X_i - m_i)^2 + \sum (X_2 - m_2)^2 + \sum (X_3 - m_3)^2$ (لكل sample)

$(1-2.67)^2 + (2-2.67)^2 + (5-2.67)^2 + (2-2.67)^2 + (4-2.67)$

$SS_{total} = \sum (X - m_G)^2 = 13.6$

لكل شيء

DATE : ____________ SUBJECT : ____________

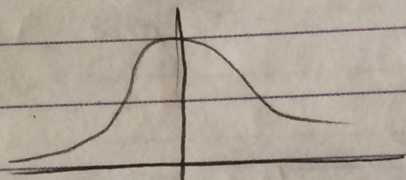
- $SS_{Between} = SS_{total} - SS_{within} = 0.23$

- $S_w^2 = V_w = \frac{SS_w}{N-K} = 13.34/(9-3) = 2.22$

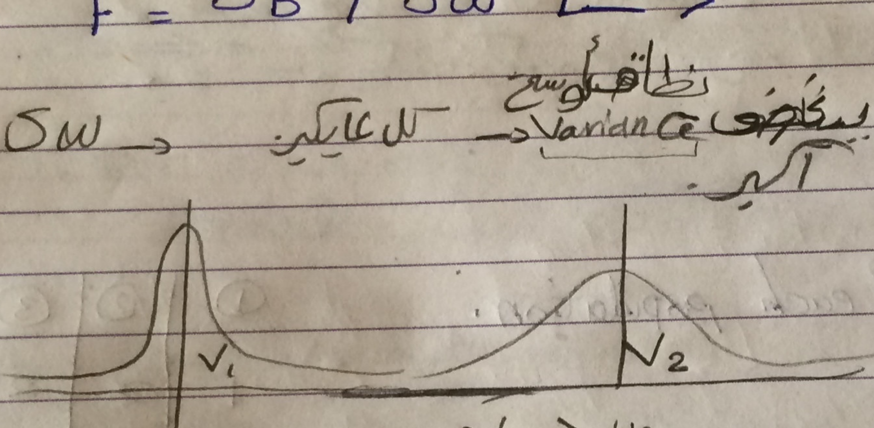
  عدد ال models "3" ← $N - K$ → عدد كل samples "9"

- $S_B^2 = \frac{SSB}{K-1} = \frac{0.23}{2} = 0.12$

- $F = S_B^2 / S_w^2$ → <1 accept null hypothes. / $>$



S_w → كل عالكيرف ← variance بيكون عرض نطاق أوسع أكبر



$V_1 > V_2$